MET O 11 TECHNICAL NOTE No. 118

A Predictor - Corrector Method and the Kalman Filter

R. DIXON

1. Introduction

Although the Kalman filter has become widely used in science and engineering since its introduction in the early sixties it is not used very much in Meteorology. The literature of it is loaded with matrix algebra and trade terms, an opaque mixture. It helps in such cases if the topic can be related to matters which are already familiar. This Note has the pedagogic purpose of pointing out the close connection between a predictor-corrector method and a version of the Kalman filter.

Met O 11 (Forecasting Research Branch)
Meteorological Office
London Road
Bracknell
Berks UK

NB This paper has not been published. Permission to quote from it must be obtained from the Assistant Director of the above Meteorological Office Branch.

2. A predictor-corrector method

Given a noisy data sequence $\{a_t\}$ we may have the real-time problem of estimating or forecasting the next value. If at time t there are available estimates $\hat{a}_t$ for the true data value and $\hat{g}_t$ for the true gradient then a forecast f_{t+1} of the next value of a is given by

$$f_{t+1} = \hat{a}_t + g_t \qquad (1)$$

To get the next forecast value f_{t+2} the estimates on the RHS of (1) have to be updated so that $\hat{a}_{t+1}$ and $\hat{g}_{t+1}$ are available. The forecast f_{t+1} will be wrong to some extent. The actual a_{t+1} will also be wrong to some extent, because of contamination by noise. The estimate $\hat{a}_{t+1}$ of the true value of a at time t+1 is therefore taken as a weighted average of a_{t+1} and f_{t+1}.

$$\hat{a}_{t+1} = \alpha a_{t+1} + (1-\alpha) f_{t+1} \qquad (2)$$

where

$$0 < \alpha < 1$$

Similar homely reasoning will provide the estimate $\hat{g}_{t+1}$. It can be taken as a weighted average of the slope from $\hat{a}_t$ to a_{t+1} and the slope from $\hat{a}_t$ to f_{t+1}, yielding

$$\hat{g}_{t+1} = \beta (a_{t+1} - \hat{a}_t) + (1-\beta)(f_{t+1} - \hat{a}_t) \qquad (3)$$

where

$$0 < \beta < 1$$

Given a choice for α and β and starting values for $\hat{a}_o$ and $\hat{g}_o$, (1), (2) and (3) provide a cyclic recursive procedure for obtaining the successive estimates and forecasts. The system (1) to (3) may be regarded as a filter having the input $\{a_t\}$ and output $\{f_t\}$ or $\{\hat{a}_t\}$. Suitable values for α and β may be arrived at by experience or from frequency domain considerations.

Suppose that we concentrate on $\{\hat{a}_t\}$ as the output from the filter. Then using Δ the standard forward difference operator, the equations (1) to (3) can be combined, eliminating f_{t+1}, to give the difference equation

$$\Delta^2 \hat{a}_{t+1} + \alpha \Delta \hat{a}_t + \beta \hat{a}_t = \alpha \Delta a_{t+1} + \beta a_t \qquad (4)$$

This can be regarded as an approximation to the differential equation

$$\frac{d^2\hat{a}}{dt^2} + \alpha \frac{d\hat{a}}{dt} + \beta \hat{a} = \alpha \frac{da}{dt} + \beta a \qquad (5)$$

The Transfer function of the system, obtained by taking the Laplace transform of (5) is

$$H(s) = \frac{\alpha s + \beta}{s^2 + \alpha s + \beta} \qquad (6)$$

Putting

$$\sqrt{\beta} = \omega_n \quad , \quad \alpha = 2\zeta\sqrt{\beta} \qquad (7)$$

(5) may then be written

$$\frac{d^2\hat{a}}{dt^2} + 2\zeta\omega_n \frac{d\hat{a}}{dt} + \omega_n^2 \hat{a} = 2\zeta\omega_n \frac{da}{dt} + \omega_n^2 a \qquad (8)$$

the equation of a forced oscillator with

ω_n the natural frequency

ζ the damping ratio

$\zeta\omega_n$ the damping coefficient

If the input a is the sinusoid of frequency

$$a = A \sin \omega t \qquad (9)$$

then the output from a system having the Transfer function (6) is $A|H(i\omega)|\sin(\omega t+\phi)$ where $\phi = arg\, H(i\omega)$, ie the ratio of the output amplitude to the input amplitude is the absolute value of $H(i\omega)$, where $H(i\omega)$ is simply (6) with s replaced by $i\omega$ $(i=\sqrt{-1})$. The bandwidth of a filter system can be described as that range of frequencies over which the filter gives a satisfactory response. What is meant by "satisfactory" will depend on the problem but as an example take the bandwidth as being defined by the frequency at which the output amplitude falls to one half the input amplitude. If this frequency is denoted ω_b then from (6) and (7)

$$\frac{1}{2} = \left| \frac{2\zeta\omega_n (i\omega_b) + \omega_n^2}{(i\omega_b)^2 + 2\zeta\omega_n (i\omega_b) + \omega_n^2} \right| \qquad (10)$$

an equation which solves to give

$$\omega_n^2 = \left[1 + 2\zeta^2 + \sqrt{2 + 4\zeta^2 + 4\zeta^4} \right] \omega_b^2 \qquad (11)$$

Having a desired bandwidth and damping ratio, ω_n can be found and thence via (7) the α and β values for the cyclic (1), (2), (3) procedure. Thus with some knowledge of the properties of the data it is possible to some extent to design the predictor-corrector $\alpha\beta$ filter to suit a given purpose.

5. An estimation theory approach to the same problem

The predictor-corrector equations of Section 2 were established by a simple commonsense argument. It is clear that the argument is not unique. It can be varied to yield slightly different versions. For example, since the data is acknowledged to be noisy it might be considered better to replace a_{t+1} by $\hat{a}_{t+1}$ in (3) to have

$$\hat{g}_{t+1} = \beta(\hat{a}_{t+1} - \hat{a}_t) + (1-\beta)(f_{t+1} - \hat{a}_t) \qquad (12)$$

It is also evident that, using (1), (12) may be written as

$$\hat{g}_{t+1} = \beta(\hat{a}_{t+1} - \hat{a}_t) + (1-\beta)\hat{g}_t \qquad (13)$$

explicitly relating $\hat{g}_{t+1}$ to the previous gradient estimate $\hat{g}_t$. Alternatively, since the data is noisy, we can argue that it is composed of a "trend" value z_t plus a noise component τ_t so that

$$a_t = z_t + \tau_t \qquad (14)$$

and that what we really need to do is to obtain an estimate $\hat{z}_t$ of the trend value and then use

$$F_{t+1} = \hat{z}_t + \hat{g}_t \qquad (15)$$

as our forecast. This leads smoothly to the stochastic model view of the problem. We assume that what we are dealing with is a "true" model such that the signal values are generated by

$$z_{t+1} = z_t + g_t \qquad (16)$$

$$g_{t+1} = g_t + \lambda_t \qquad (17)$$

where λ_t is an error term. The values we actually observe are the values given by (14), and the recursive estimation scheme is now

$$\hat{z}_{t+1} = \alpha a_{t+1} + (1-\alpha)F_{t+1} \qquad (18)$$

$$\hat{g}_{t+1} = \beta(\hat{z}_{t+1} - \hat{z}_t) + (1-\beta)\hat{g}_t \qquad (19)$$

The set of equations (15), (18) and (19) is another reasonable variant of the set (1), (2), (3). Appropriate values for α and β can still be chosen from experience or by the frequency domain considerations of Section 2, but more in keeping with the spirit of this Section, assuming that a set of data is already available, subject to the conditions

$$E(\underline{\lambda}) = E(\underline{\tau}) = \underline{0} \qquad (20)$$

$$E(\underline{\lambda}\,\underline{\tilde{\lambda}}) = \sigma_\lambda^2 I \qquad (21)$$

$$E(\underline{\tau}\,\underline{\tilde{\tau}}) = \sigma_\tau^2 I \qquad (22)$$

$$E(\underline{\lambda}\,\underline{\tilde{\tau}}) = 0 \quad \text{(zero matrix)} \qquad (23)$$

it can be shown that the values of α and β which minimize the prediction error, in the least-squares sense, are

$$\overset{\downarrow}{\alpha} = \frac{2\ell}{1+\ell} \qquad (24)$$

$$\overset{\downarrow}{\beta} = \ell \qquad (25)$$

where

$$\ell^2 = \frac{1}{8}\,\frac{\sigma_\lambda}{\sigma_+}\left[\sqrt{\frac{\sigma_\lambda^2}{\sigma_+^2}+16} - 1\right] \qquad (26)$$

These "optimal" values of $\overset{\vee}{\alpha}$ and $\overset{\vee}{\beta}$ could be found at each time-step either for some fixed length of series or for a lengthening series, a cumbersome but possible procedure.

4. The Kalman filter approach

In Met O 11 Technical Note No. 112 the state-space and Kalman filter equations of a time dependent system are given, equations (2) to (9). For definitions, notation, and an elementary explanation see Sections 2 and 3 of Note 112. If we deal with the slightly more general system

$$\underline{a}_t = M_{t-1} \cdot \underline{a}_{t-1} + B_{t-1} \cdot \underline{\lambda}_{t-1} \qquad (27)$$

$$\underline{h}_t = F_t \cdot \underline{a}_t + \underline{r}_t \qquad (28)$$

then the Kalman filter equations become

$$\overset{*}{S}_t = M_{t-1} \cdot \hat{S}_{t-1} \cdot \tilde{M}_{t-1} + B_{t-1} \cdot \Lambda_{t-1} \cdot \tilde{B}_{t-1} \qquad (29)$$

$$\hat{S}_t^{-1} = \overset{*}{S}_t^{-1} + \tilde{F}_t \cdot R_t^{-1} \cdot F_t \qquad (30)$$

$$K_t = \hat{S}_t \cdot F_t \cdot R_t^{-1} \qquad (31)$$

$$\overset{*}{\underline{a}}_t = M_{t-1} \cdot \hat{\underline{a}}_{t-1} \qquad (32)$$

$$\hat{\underline{a}}_t = \overset{*}{\underline{a}}_t + K_t \cdot (\underline{h}_t - F_t \cdot \overset{*}{\underline{a}}_t) \qquad (33)$$

where S_{t-1} is the covariance matrix

$$S_{t-1} = E\,(\hat{\underline{a}}_{t-1} - \underline{a}_{t-1})(\hat{\underline{a}}_{t-1} - \underline{a}_{t-1}) \qquad (34)$$

Looking now at (16), (17) and (14) of Section 3 of this Note it is seen that they can be written in matrix/vector notation as

$$\begin{pmatrix} z_{t+1} \\ g_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} z_t \\ g_t \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \lambda_t \qquad (35)$$

and

$$a_t = (1 \quad 0) \cdot \begin{pmatrix} z_t \\ g_t \end{pmatrix} + +_t \qquad (36)$$

and it is now apparent the model of Section 3 above is a particular case of the more general system (27) and (28) with

$$\left.\begin{aligned} \underline{a}_t &= \begin{pmatrix} z_t \\ g_t \end{pmatrix}, & M_{t-1} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ B_{t-1} &= \begin{pmatrix} 0 \\ 1 \end{pmatrix}, & F_t &= (1 \quad 0) \\ \Lambda_{t-1} &= \sigma_\lambda^2 I, & R_t &= \sigma_r^2 I \end{aligned}\right\} \qquad (37)$$

Thus with these specializations (32) yields

$$\begin{pmatrix} \overset{*}{z}_{t+1} \\ \overset{*}{g}_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{z}_t \\ \hat{g}_t \end{pmatrix} \qquad (38)$$

and (33) yields

$$\begin{pmatrix} \hat{z}_{t+1} \\ \hat{g}_{t+1} \end{pmatrix} = \begin{pmatrix} \overset{*}{z}_{t+1} \\ \overset{*}{g}_{t+1} \end{pmatrix} + \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}_{t+1} \cdot \left\{ a_{t+1} - (1 \quad 0) \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{z}_t \\ \hat{g}_t \end{pmatrix} \right\} \qquad (39)$$

Putting (38) into (39) there results

$$\begin{pmatrix} \hat{z}_{t+1} \\ \hat{g}_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{z}_t \\ \hat{g}_t \end{pmatrix} + \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \cdot \left\{ a_{t+1} - (1 \quad 0) \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{z}_t \\ \hat{g}_t \end{pmatrix} \right\} \qquad (40)$$

where the gain vector $(k_1 \quad k_2)\sim$ is given by (31), taking account of (37), as

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix}_{t+1} = \hat{S}_{t+1} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \sigma_r^2 \qquad (41)$$

The interest of (40) is that in Section 3 by combining (15), (18) and (19) they may be put into the matrix/vector form

$$\begin{pmatrix} \hat{z}_{t+1} \\ \hat{g}_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{z}_t \\ \hat{g}_t \end{pmatrix} + \begin{pmatrix} \alpha \\ \alpha\beta \end{pmatrix} \cdot \left\{ a_{t+1} - (1 \quad 0) \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \hat{z}_t \\ \hat{g}_t \end{pmatrix} \right\} \qquad (42)$$

and comparing (40) with (42) it is seen that they are identical in form. As customarily used, the $\alpha\beta$ filter will not give the same results as the Kalman filter because α and β are generally fixed a priori whereas the components k_1 and k_2 of the Kalman gain vector change at each time-step. However, if the procedure in the final sentence of Section 3 is being followed and an optimal least-squares $\overset{\Psi}{\alpha}$ and $\overset{\Psi}{\beta}$ is calculated at each time-step for a series which is lengthening by one at each time-step then $\overset{\Psi}{\alpha}$ and $\overset{\Psi}{\beta}$ change each time-step and it can then be shown that it is possible to arrange for

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix}_t = \begin{pmatrix} \overset{\Psi}{\alpha} \\ \overset{\Psi}{\alpha}\ \overset{\Psi}{\beta} \end{pmatrix}_t \qquad (43)$$

for all t. It can be seen from (41) that this will only be possible if S_0 the starting value of S_t, is suitably chosen. It is found that the correct value of S_0 for (43) to be true at all t is

$$S_0 = \frac{2l\sigma_+^2}{1+l} \begin{pmatrix} 1 & l \\ l & \frac{2l^2}{1-l} \end{pmatrix} \qquad (44)$$

l and σ_+^2 being as defined in Section 3.

5. Concluding remarks

By drawing together certain strands and achieving a measure of unification it is hoped that this Note brings some enlightenment on a topic of current interest. It should be appreciated that none of the procedures shown above to be stongly connected are in any case God-given. They are all man-made and as such can doubtless be improved upon and troubled by counter-examples.

R Dixon

R Dixon
Met O 11

October 1978